Lecture 20: Omitted Variable Bias

Prof. Esther Duflo

14.310x

Non linear transformation of the independent variables

- When running a kernel regression as exploratory analysis we may realize that the relationship between two variables does not appear to be linear.
- Does it mean we cannot run OLS?

Non linear transformation of the independent variables

- When running a kernel regression as exploratory analysis we may realize that the relationship between two variables does not appear to be linear.
- Does it mean we cannot run OLS?
- No!
- We can use polynomial or other transformations of the data to represent non linearities
- or partition the range of X.

Polynomial models

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{1i}^2 + \cdots + \beta_k X_{1i}^k + E_i$$

- You can chose straight polynomial, or series expansion, or orthogonal polynomials or whatever.
- If you assume that the model is known, this is just standard OLS. You may want to plot the curve, or compute the derivative with respect to X at key points, etc.
- If you assume that the model is not known, this is a non-parametric method: you realize there is bias (because the shape is never quite perfect) and variance (as you add more Xs) and you promise to add more terms as the number of observation increases. This is called *series* regression.

Other non linear transformations

- Take log of X
- Interact the X, such as the slope of one depends on the level of another.
- Potentially lots of variables and their transformations... How to chose? This is where machine learning tools can become handy (more on that later!)

Using dummies for approximation

- Partition the range of X is interval, $X_0, \ldots X_J$
- Define the dummies as:

$$
\begin{aligned}
D_{1i} &= I_{[X_0 \leq X_{1i} < X_1]} \\
D_{2i} &= I_{[X_1 \leq X_{1i} < X_2]} \\
&\vdots \\
D_{ji} &= I_{[X_{J-1} \leq X_{1i} < X_J]}
\end{aligned}
$$

Using dummies for approximation

- Partition the range of X is interval, $X_0, \ldots X_J$
- Define the dummies as:

$$\begin{aligned} D_{1i} &= I_{[X_0 \leq X_{1i} < X_1]} \\ D_{2i} &= I_{[X_1 \leq X_{1i} < X_2]} \\ &\vdots \\ D_{ji} &= I_{[X_{J-1} \leq X_{1i} < X_J]} \end{aligned}$$

- you can run regression:

$$Y_i = \beta_1 D_{1i} + \beta_2 D_{2i} + \cdots + \beta_J D_{ji} + E_i$$

(note no intercept. why?)

Using dummies for approximation

- Partition the range of X is interval, $X_0, \ldots X_J$
- Define the dummies as:

$$\begin{aligned} D_{1i} &= I_{[X_0 \leq X_{1i} < X_1]} \\ D_{2i} &= I_{[X_1 \leq X_{1i} < X_2]} \\ &\vdots \\ D_{ji} &= I_{[X_{J-1} \leq X_{1i} < X_J]} \end{aligned}$$

- you can run regression:

$$Y_i = \beta_1 D_{1i} + \beta_2 D_{2i} + \cdots + \beta_J D_{ji} + E_i$$

(note no intercept. why?)

- Define Piece wise linear variables as:

$S_{1i} = I_{[X_0 \leq X_{1i} < X_1]}(X_{1i} - X_1)$ $S_{2i} = I_{[X_1 \leq X_{1i} < X_2]}(X_{1i} - X_2)$

Using dummies for approximation

- Partition the range of X is interval, $X_0, \ldots X_J$
- Define the dummies as:

$$
\begin{aligned}
D_{1i} &= I_{[X_0 \leq X_{1i} < X_1]} \\
D_{2i} &= I_{[X_1 \leq X_{1i} < X_2]} \\
&\vdots \\
D_{ji} &= I_{[X_{J-1} \leq X_{1i} < X_J]}
\end{aligned}
$$

- you can run regression:

$$Y_i = \beta_1 D_{1i} + \beta_2 D_{2i} + \cdots + \beta_J D_{ji} + E_i$$

(note no intercept. why?)

- Define Piece wise linear variables as:

$S_{1i} = I_{[X_0 \leq X_{1i} < X_1]}(X_{1i} - X_1)$ $S_{2i} = I_{[X_1 \leq X_{1i} < X_2]}(X_{1i} - X_2)$

- Run regression

$$Y_i = \beta_1 X_{1i} + \beta_2 S_{1i} + \cdots + \beta_J S_{j-1i} + E_i$$

Locally Linear Regression

- What size of interval should we chose?
- This should by now sound very familiar: either you are willing to assume that you *know* the shape of the function: Then, just cut it as you know it is relevant.
- Or.... we are trying to guess the shape of the function
- And then we have the familiar bias/variance trade off: we are now in fact performing a non parametric regression technique known as a locally linear regression: around each point where we are interested in evaluating the function, we run a weighted regression of Y_i on X_i, where the weights will be given by a Kernel, for the set of observations within the bandwidth. We take the predicted value from the regression as best predictor for Y_i. So it is exactly like a Kernel regression, but we use a linear regression in each little interval instead!
- Why on earth?

Locally Linear Regression

- What size of interval should we chose?
- This should by now sound very familiar: either you are willing to assume that you *know* the shape of the function: Then, just cut it as you know it is relevant.
- Or.... we are trying to guess the shape of the function
- And then we have the familiar bias/variance trade off: we are now in fact performing a non parametric regression technique known as a locally linear regression: around each point where we are interested in evaluating the function, we run a weighted regression of Y_i on X_i, where the weights will be given by a Kernel, for the set of observations within the bandwidth. We take the predicted value from the regression as best predictor for Y_i. So it is exactly like a Kernel regression, but we use a linear regression in each little interval instead!
- Why on earth?
    - It has better properties (especially at the boundaries)
    - And the slope is often of interest

Putting this all together: Regression Discontinuity Design

- One application of all these methods together is the popular “regression discontinuity design” to evaluate causal effects.
- RD is appropriate in any circumstance where some treatment shift discontinuously with a variable a, called the *running variable*
- E.g. a scholarship attributed to those with at least P points; an election won or lost at 50%.
- E.g. $D_a = 1$ if $a >= 21$ and 0 otherwise. [guess what is D_a?]
- The idea is that the outcome Y_i may change with the running variable, but we assume that it would not change discontinuously at some threshold a_0, if it did not force the first stage.

Analysis

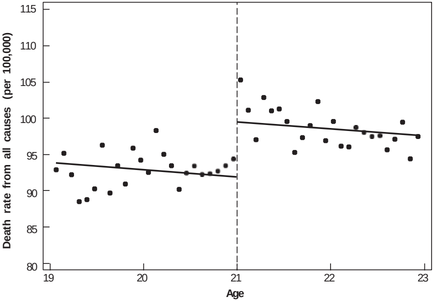
Simplest analysis: Use the dummy variable to shift the intercept at a_0

$$Y_i = \beta_0 + \beta_1 D_{ai} + \beta_2 a_i + \mathcal{E}_i$$

where Y_i is road fatalities, and D_{ai} is dummy for being allowed to drink and a is age.

Minimum drinking age

Figure 4.2
A sharp RD estimate of MLDA mortality effects

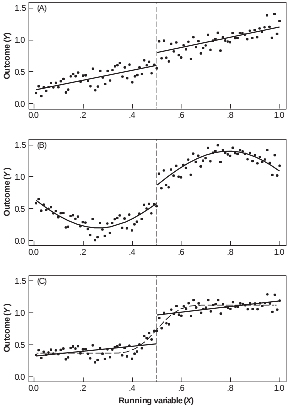


Notes: This figure plots death rates from all causes against age in months. The lines in the figure show fitted values from a regression of death rates on an over-21 dummy and age in months (the vertical dashed line indicates the minimum legal drinking age (MLDA) cutoff).

Source: Angrist and Pischke ”Mastering Metrics”, Figure 4.2
Original Data from: “The Effect of Alcohol Access on Consumption and Mortality: Regression Discontinuity Evidence from the Minimum Drinking Age,” with Christopher Carpenter, American Economic Journal: Applied Economics, Vol. 1, Issue 1, pp. 164-82

However the simplest analysis may get it wrong... in particular non linearities may disguise themselves as discontinuities!

Figure 4.3
RD in action, three ways

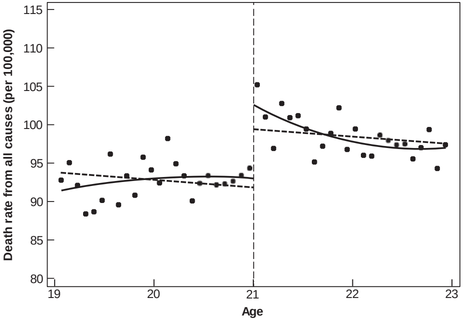


Notes: Panel A shows RD with a linear model for $E[Y_i|X_i]$; panel B adds some curvature. Panel C shows nonlinearity mistaken for a discontinuity. The vertical dashed line indicates a hypothetical RD cutoff.

Source: Angrist and Pischke "Mastering Metrics", Figure 4.3

How do we solve this problem? (1)

Figure 4.4
Quadratic control in an RD design



Notes: This figure plots death rates from all causes against age in months. Dashed lines in the figure show fitted values from a regression of death rates on an over-21 dummy and age in months. The solid lines plot fitted values from a regression of mortality on an over-21 dummy and a quadratic in age, interacted with the over-21 dummy (the vertical dashed line indicates the minimum legal drinking age [MLDA] cutoff).

source: Angrist and Pischke ”Mastering Metrics”, Figure 4.4

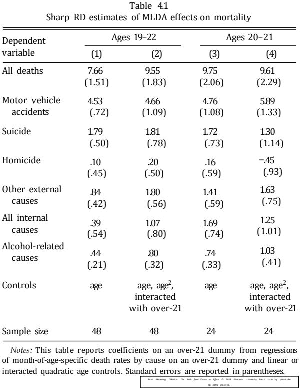


Table 4.1
Sharp RD estimates of MLDA effects on mortality

Dependent variable	Ages 19–22		Ages 20–21	
	(1)	(2)	(3)	(4)
All deaths	7.66 (1.51)	9.55 (1.83)	9.75 (2.06)	9.61 (2.29)
Motor vehicle accidents	4.53 (.72)	4.66 (1.09)	4.76 (1.08)	5.89 (1.33)
Suicide	1.79 (.50)	1.81 (.78)	1.72 (.73)	1.30 (1.14)
Homicide	.10 (.45)	.20 (.50)	.16 (.59)	−.45 (.93)
Other external causes	.84 (.42)	1.80 (.56)	1.41 (.59)	1.63 (.75)
All internal causes	.39 (.54)	1.07 (.80)	1.69 (.74)	1.25 (1.01)
Alcohol-related causes	.44 (.21)	.80 (.32)	.74 (.33)	1.03 (.41)
Controls	age	age, age^2, interacted with over-21	age	age, age^2, interacted with over-21
Sample size	48	48	24	24

Notes: This table reports coefficients on an over-21 dummy from regressions of month-of-age-specific death rates by cause on an over-21 dummy and linear or interacted quadratic age controls. Standard errors are reported in parentheses.

source: Angrist and Pischke ”Mastering Metrics”, Table 4.1

How do we solve this problem? (2)

We can also solve the problem by narrowing the estimate to a band around the discontinuity (the bandwidth!). As usual, the risk is bias vs variance: if we promise to narrow the bandwidth as the number of observation increases, we now have a non -parametric RD!

The Omitted Variables Bias: An example

Imagine that you are interested in estimating the impact of going to a private college (vs a state school) on earnings. This example comes from Dale and Krueger, via Angrist and Pishke “Masters of Metrics” textbook (chapter 2). The data focuses on people who enrolled in college in 1976. It has their SAT score, their parental income, where they went to college, and where they applied and were admitted.

The true model is

$$\ln(Y_i) = \alpha + \beta P_i + \sum_{j=1}^{150} \gamma_j Group_{ij} + \delta_1 SAT_i + \delta_2 PI_i + E_i$$

where $\ln(Y_i)$ is log earnings later in life; SAT_i is the SAT score of student i and P_i their parental income, and some other demographic characteristics, and $Group_{ij}$ is defined in the next slide.

Capturing application behavior

$Group_{ij}$ is a dummy equal to 1 if student *i* belongs to group *j* (a set of "group" fixed effects, as we saw in the previous lecture - I had used the notation α_j to indicate group fixed effects). These groups describe the set of schools the students have applied to and where they were admitted (e.g. all the students that have applied to 3 selective schools and 1 non selective school, and got admitted to all the places they applied to are in one group).

We are interested in β. Why do all the other variables belong to the full model?

Capturing application behavior

Group$_{ij}$ is a dummy equal to 1 if student *i* belongs to group *j* (a set of "group" fixed effects, as we saw in the previous lecture - I had used the notation α_j to indicate group fixed effects). These groups describe the set of schools the students have applied to and where they were admitted (e.g. all the students that have applied to 3 selective schools and 1 non selective school, and got admitted to all the places they applied to are in one group).

We are interested in β. Why do all the other variables belong to the full model?

- They are controlling for **selection bias:** by writing down this model, we are assuming that, once we have accounted for these variables, the potential outcomes would have been the same for those who attended a private college and those who did not.

Capturing application behavior

Group$_{ij}$ is a dummy equal to 1 if student *i* belongs to group *j* (a set of "group" fixed effects, as we saw in the previous lecture - I had used the notation α_j to indicate group fixed effects). These groups describe the set of schools the students have applied to and where they were admitted (e.g. all the students that have applied to 3 selective schools and 1 non selective school, and got admitted to all the places they applied to are in one group).

We are interested in β. Why do all the other variables belong to the full model?

- They are controlling for **selection bias:** by writing down this model, we are assuming that, once we have accounted for these variables, the potential outcomes would have been the same for those who attended a private college and those who did not.
- Why do we include SAT score and parental income in the model?

Capturing application behavior

Group$_{ij}$ is a dummy equal to 1 if student *i* belongs to group *j* (a set of "group" fixed effects, as we saw in the previous lecture - I had used the notation α_j to indicate group fixed effects). These groups describe the set of schools the students have applied to and where they were admitted (e.g. all the students that have applied to 3 selective schools and 1 non selective school, and got admitted to all the places they applied to are in one group).

We are interested in β. Why do all the other variables belong to the full model?

- They are controlling for **selection bias:** by writing down this model, we are assuming that, once we have accounted for these variables, the potential outcomes would have been the same for those who attended a private college and those who did not.
- Why do we include SAT score and parental income in the model?
- Why do we include "group" fixed effects for the set of schools other people applied to and were admitted to?

Now imagine that you don’t have all these variables, so you run partial models:

No controls:

$$Y_i = \alpha + \beta P_i + E_i$$

Control just for SAT :

$$Y_i = \alpha + \beta P_i + \delta_1 SAT_i + E_i$$

Control just for SAT and other characteristics:

$$Y_i = \alpha + \beta P_i + \delta_1 SAT_i + +\delta_2 PI_i + E_i$$

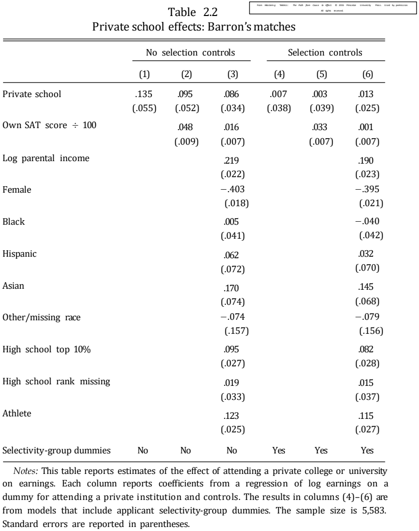


Table 2.2
Private school effects: Barron's matches

	No selection controls			Selection controls		
	(1)	(2)	(3)	(4)	(5)	(6)
Private school	.135 (.055)	.095 (.052)	.086 (.034)	.007 (.038)	.003 (.039)	.013 (.025)
Own SAT score ÷ 100		.048 (.009)	.016 (.007)		.033 (.007)	.001 (.007)
Log parental income			.219 (.022)			.190 (.023)
Female			−.403 (.018)			−.395 (.021)
Black			.005 (.041)			−.040 (.042)
Hispanic			.062 (.072)			.032 (.070)
Asian			.170 (.074)			.145 (.068)
Other/missing race			−.074 (.157)			−.079 (.156)
High school top 10%			.095 (.027)			.082 (.028)
High school rank missing			.019 (.033)			.015 (.037)
Athlete			.123 (.025)			.115 (.027)
Selectivity-group dummies	No	No	No	Yes	Yes	Yes

Notes: This table reports estimates of the effect of attending a private college or university on earnings. Each column reports coefficients from a regression of log earnings on a dummy for attending a private institution and controls. The results in columns (4)–(6) are from models that include applicant selectivity-group dummies. The sample size is 5,583. Standard errors are reported in parentheses.

- You can see that adding the SAT and demographic controls reduces the private school “premium” somewhat
- Controlling for the group dummies reduces the “premium” to zero.
- What is happening?

- You can see that adding the SAT and demographic controls reduces the private school “premium” somewhat
- Controlling for the group dummies reduces the “premium” to zero.
- What is happening?
- (A note: we are also losing a lot of observations when we do this, but if instead you control for the number of schools people applied to, we get very similar numbers–see next table)

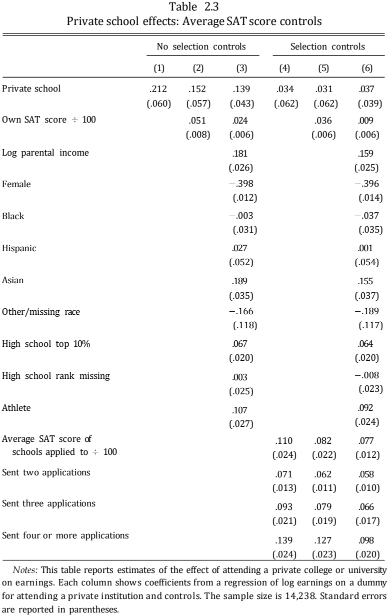


Table 2.3
Private school effects: Average SAT score controls

	No selection controls			Selection controls		
	(1)	(2)	(3)	(4)	(5)	(6)
Private school	.212 (.060)	.152 (.057)	.139 (.043)	.034 (.062)	.031 (.062)	.037 (.039)
Own SAT score ÷ 100		.051 (.008)	.024 (.006)		.036 (.006)	.009 (.006)
Log parental income			.181 (.026)			.159 (.025)
Female			−.398 (.012)			−.396 (.014)
Black			−.003 (.031)			−.037 (.035)
Hispanic			.027 (.052)			.001 (.054)
Asian			.189 (.035)			.155 (.037)
Other/missing race			−.166 (.118)			−.189 (.117)
High school top 10%			.067 (.020)			.064 (.020)
High school rank missing			.003 (.025)			−.008 (.023)
Athlete			.107 (.027)			.092 (.024)
Average SAT score of schools applied to ÷ 100				.110 (.024)	.082 (.022)	.077 (.012)
Sent two applications				.071 (.013)	.062 (.011)	.058 (.010)
Sent three applications				.093 (.021)	.079 (.019)	.066 (.017)
Sent four or more applications				.139 (.024)	.127 (.023)	.098 (.020)

Notes: This table reports estimates of the effect of attending a private college or university on earnings. Each column shows coefficients from a regression of log earnings on a dummy for attending a private institution and controls. The sample size is 14,238. Standard errors are reported in parentheses.

The omitted variable Bias formula

Correct model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + E_i$$

Estimated model:

$$Y_i = \alpha_0 + \alpha_1 X_{1i} + w_i$$

Define Ancillary (or Auxillary) regression as:

$$X_{2i} = \delta_0 + \delta_1 X_{1i} + \xi_i$$

The omitted variable Bias formula

Correct model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + E_i$$

Estimated model:

$$Y_i = \alpha_0 + \alpha_1 X_{1i} + w_i$$

Define Ancillary (or Auxillary) regression as:

$$X_{2i} = \delta_0 + \delta_1 X_{1i} + \xi_i$$

Then:

$$OVB = \alpha_1 - \beta_1 = \delta_1 \beta_2$$

The omitted variable Bias formula

Correct model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + E_i$$

Estimated model:

$$Y_i = \alpha_0 + \alpha_1 X_{1i} + w_i$$

Define Ancillary (or Auxillary) regression as:

$$X_{2i} = \delta_0 + \delta_1 X_{1i} + \xi_i$$

Then:

$$OVB = \alpha_1 - \beta_1 = \delta_1 \beta_2$$

Wow! isn't that nifty!!!!

Analyzing the OVB formula

Omitted variable bias depends on :

1. How important is X_2 in the original model
2. How correlated it is with X_1

What is the intuition for this result ?

Bivariate derivation

It is worth spending some time with this formula because it is going to stay with you for your entire life as a data scientist! Remember OLS bivariate formula:

$$\alpha_1 = \frac{Cov(Y_i, X_{1i})}{V(X_{1i})}$$

Bivariate derivation

It is worth spending some time with this formula because it is going to stay with you for your entire life as a data scientist! Remember OLS bivariate formula:

$$\alpha_1 = \frac{Cov(Y_i, X_{1i})}{V(X_{1i})}$$

substituting for Y_i we get:

$$\frac{Cov(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + E_i, X_{1i})}{V(X_{1i})}$$

Bivariate derivation

It is worth spending some time with this formula because it is going to stay with you for your entire life as a data scientist! Remember OLS bivariate formula:

$$\alpha_1 = \frac{Cov(Y_i, X_{1i})}{V(X_{1i})}$$

substituting for Y_i we get:

$$\frac{Cov(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + E_i, X_{1i})}{V(X_{1i})}$$

$$= \frac{\beta_1 V(X_{1i}) + \beta_2 Cov(X_{2i}, X_{1i}) + Cov(E_i, X_{1i})}{V(X_{1i})}$$

Bivariate derivation

It is worth spending some time with this formula because it is going to stay with you for your entire life as a data scientist! Remember OLS bivariate formula:

$$\alpha_1 = \frac{Cov(Y_i, X_{1i})}{V(X_{1i})}$$

substituting for Y_i we get:

$$\frac{Cov(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + E_i, X_{1i})}{V(X_{1i})}$$

$$= \frac{\beta_1 V(X_{1i}) + \beta_2 Cov(X_{2i}, X_{1i}) + Cov(E_i, X_{1i})}{V(X_{1i})}$$

$$= \beta_1 + \delta_1 \beta_2$$

Matrix Proof

$$\alpha_1 = (X_1^t X_1)^{-1} X_1^t Y = (X_1^t X_1)^{-1} X_1^t(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + E)$$

$$= \beta_1 + (X_1^t X_1)^{-1} X_1^t X_2 \beta_2 + (X_1^t X_1)^{-1} X_1^t E$$

Matrix Proof

$$\alpha_1 = (X_1^t X_1)^{-1} X_1^t Y = (X_1^t X_1)^{-1} X_1^t (\beta_0 + \beta_1 X_1 + \beta_2 X_2 + E)$$

$$= \beta_1 + (X_1^t X_1)^{-1} X_1^t X_2 \beta_2 + (X_1^t X_1)^{-1} X_1^t E$$

$$= \beta_1 + (X_1^t X_1)^{-1} X_1^t X_2 \beta_2 = \beta_1 + \delta_1 \beta_2$$

Back to our college example

For illustration, suppose that the true model only included SAT score as a control, how does the OVB formula work?

We need to run the **auxiliary** regression :

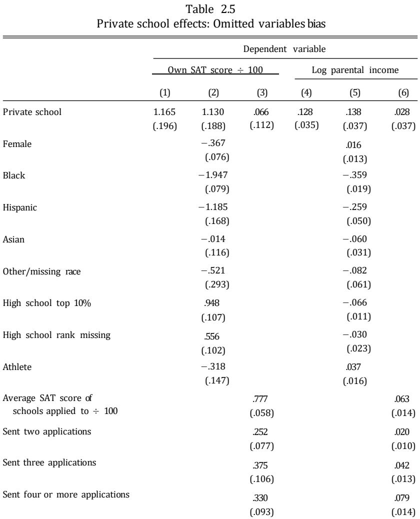


Table 2.5
Private school effects: Omitted variables bias

	Dependent variable					
	Own SAT score ÷ 100			Log parental income		
	(1)	(2)	(3)	(4)	(5)	(6)
Private school	1.165	1.130	.066	.128	.138	.028
	(.196)	(.188)	(.112)	(.035)	(.037)	(.037)
Female		−.367			.016	
		(.076)			(.013)	
Black		−1.947			−.359	
		(.079)			(.019)	
Hispanic		−1.185			−.259	
		(.168)			(.050)	
Asian		−.014			−.060	
		(.116)			(.031)	
Other/missing race		−.521			−.082	
		(.293)			(.061)	
High school top 10%		.948			−.066	
		(.107)			(.011)	
High school rank missing		.556			−.030	
		(.102)			(.023)	
Athlete		−.318			.037	
		(.147)			(.016)	
Average SAT score of schools applied to ÷ 100			.777			.063
			(.058)			(.014)
Sent two applications			.252			.020
			(.077)			(.010)
Sent three applications			.375			.042
			(.106)			(.013)
Sent four or more applications			.330			.079
			(.093)			(.014)

Recall previous results

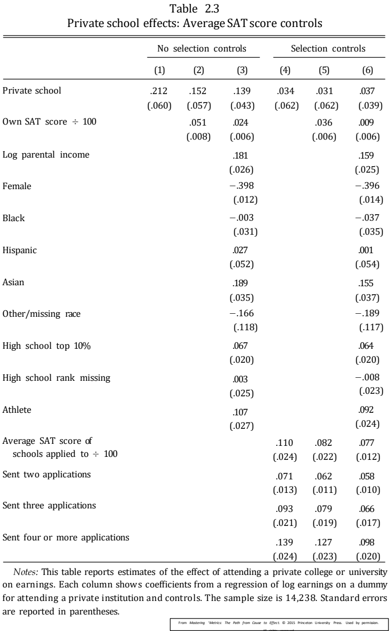


Table 2.3
Private school effects: Average SAT score controls

	No selection controls			Selection controls		
	(1)	(2)	(3)	(4)	(5)	(6)
Private school	.212	.152	.139	.034	.031	.037
	(.060)	(.057)	(.043)	(.062)	(.062)	(.039)
Own SAT score ÷ 100		.051	.024		.036	.009
		(.008)	(.006)		(.006)	(.006)
Log parental income			.181			.159
			(.026)			(.025)
Female			−.398			−.396
			(.012)			(.014)
Black			−.003			−.037
			(.031)			(.035)
Hispanic			.027			.001
			(.052)			(.054)
Asian			.189			.155
			(.035)			(.037)
Other/missing race			−.166			−.189
			(.118)			(.117)
High school top 10%			.067			.064
			(.020)			(.020)
High school rank missing			.003			−.008
			(.025)			(.023)
Athlete			.107			.092
			(.027)			(.024)
Average SAT score of schools applied to ÷ 100				.110	.082	.077
				(.024)	(.022)	(.012)
Sent two applications				.071	.062	.058
				(.013)	(.011)	(.010)
Sent three applications				.093	.079	.066
				(.021)	(.019)	(.017)
Sent four or more applications				.139	.127	.098
				(.024)	(.023)	(.020)

Notes: This table reports estimates of the effect of attending a private college or university on earnings. Each column shows coefficients from a regression of log earnings on a dummy for attending a private institution and controls. The sample size is 14,238. Standard errors are reported in parentheses.

Checking the formula

OVB (Short minus Long)=0.212-0.152=0.06

$OVB = \delta_1\beta_2 = 1.165 * 0.051 = 0.06$ (Yeah!)

Understanding why the applications dummies help

- Now suppose the “true” model included application dummies and SAT

Understanding why the applications dummies help

- Now suppose the “true” model included application dummies and SAT
- And we estimate it ignoring the SAT but including the application dummies.
- Once we control for the group application dummies, SAT score is not really correlated with private school attendance.
- So the omitted variable bias of not including it is low.
- What could be going on?

Understanding why the applications dummies help

- Now suppose the "true" model included application dummies and SAT
- And we estimate it ignoring the SAT but including the application dummies.
- Once we control for the group application dummies, SAT score is not really correlated with private school attendance.
- So the omitted variable bias of not including it is low.
- What could be going on?
- We sometimes call exercises like table 2.2 "regression sensitivity analysis": it gives us *some* confidence that the set of controls we do have reasonably captures heterogeneity between people. But of course we cannot prove it.

How do we use the OVB formula in general?

- Most of the times we don't have the variables we are not including... otherwise we would include them!
- So how is the OVB formula useful?

How do we use the OVB formula in general?

- Most of the times we don't have the variables we are not including... otherwise we would include them!
- So how is the OVB formula useful?
- It guides our economic thinking on whether the bias would be important

How do we use the OVB formula in general?

- Most of the times we don't have the variables we are not including... otherwise we would include them!
- So how is the OVB formula useful?
- It guides our economic thinking on whether the bias would be important
  - When we are running a regression, are we omitting variables that are likely to be important determinant of the *outcome*
  - And are they likely to be correlated with the *regressor of interest*

Describes the following regression:

$$SL_i = \beta_0 + \beta_1 HF_i + \beta_2 X_{2i} + E_i$$

The variables they include try to control for things that may be correlated with *HF* and affect diet, but what if we had also included some measure for self-control? Or a more continuous measure of exercise? Or what else?
It may well be true that high fat diet affects daytime sleepiness but it would be incorrect to conclude that you'd obtain the same results in an experimental setting ...

Some hints of more advanced techniques...

- Matching techniques: control as flexibly as possible for a (fixed) set of covariates which are known to be correlated with treatment.
    - Control for each group dummies (when variable are categorical).
    - Control flexibly for “propensity score’ =predicted probability to be treated, based on non parametric regression on the covariates we have.

Some hints of more advanced techniques...

- Machine Learning techniques: Double Post LASSO (Chernozoukov and Hansen)
    - Suppose that that we have lots of variables and we are not sure what variable we *should* include.
    - There are machine learning techniques to learn which variables are "predictive", i.e. to "fish" which variables should enter in a regression (we will talk more about this starting wednesday), one of which is LASSO
    - C and H propose a conceptually very simple technique in 3 steps:
        1. Regress X_1 on all the available variables, and see what LASSO picks. Call this X_2
        2. Regress Y on all the available variables, and see what LASSO picks. X_3
        3. Runs $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + E_i$

Conclusion

- Because we cannot run experiments for everything, we often attempt to get at causality by controlling for variables we can observe.
- Sometimes we can get quite close, but we will always have to make the argument that we have controlled for everything we can
- The omitted variable bias helps us think through what bias may still remain
- And sometimes we won't be willing to do this! This is when we need to use other econometric techniques... [or give up and run an experiment :-)]

References

Angrist and Pishke Masters of Metrics, Chapter 2